Wikification and Beyond: The Challenges of Entity and Concept Grounding

Dan Roth
University of Illinois at Urbana-Champaign
danr@illinois.edu

Heng Ji
Rensselaer Polytechnic Institute
jih@rpi.edu

Ming-Wei Chang
Microsoft Research
minchang@microsoft.com

Taylor Cassidy
Army Research Lab & IBM Research
taylor.cassidy.ctr@mail.mil

1 Introduction

Contextual disambiguation and grounding of concepts and entities in natural language are essential to progress in many natural language understanding tasks and fundamental to many applications. Wikification aims at automatically identifying concept mentions in text and linking them to referents in a knowledge base (KB) (e.g., Wikipedia). Consider the sentence, "*The **Times** report on **Blumenthal (D)** has the potential to fundamentally reshape the contest in **the Nutmeg State***.". A Wikifier should identify the key entities and concepts and map them to an encyclopedic resource (e.g., "*D*" refers to *Democratic Party*, and "*the Nutmeg State*" refers to *Connecticut*.

Wikification benefits end-users and Natural Language Processing (NLP) systems. Readers can better comprehend Wikified documents as information about related topics is readily accessible. For systems, a Wikified document elucidates concepts and entities by grounding them in an encyclopedic resource or an ontology. Wikification output has improved NLP down-stream tasks, including coreference resolution, user interest discovery , recommendation and search.

This task has received increased attention in recent years from the NLP and Data Mining communities, partly fostered by the U.S. NIST Text Analysis Conference Knowledge Base Population (KBP) track, and several versions of it has been studied. These include Wikifying all concept mentions in a single text document; Wikifying a cluster of co-referential named entity mentions that appear across documents (Entity Linking), and Wikifying a whole document to a single concept. Other works relate this task to coreference resolution within and across documents and in the context of multiple text genres.

2 Content Overview

This tutorial will motivate Wikification as a broad paradigm for cross-source linking for knowledge enrichment. We will discuss multiple dimensions of the task definition, present the building blocks of a state-of-the-art Wikifier, share key lessons learned from analysis of results, and discuss recently proposed ideas for advancing work in this area in response to key challenges. We will touch on new research areas including interactive Wikification, social media, and censorship. The tutorial will be useful for all those with interests in cross-source information extraction and linking, knowledge acquisition, and the use of acquired knowledge in NLP. We will provide a concise roadmap of recent perspectives and results, and point to some of our available Wikification resources.

3 Outline

- Introduction and Motivation
- Methodological presentation of a skeletal Wikification system
  - Mention and candidate identification
  - Knowledge representation
  - Local and global context analysis
  - Role of Machine Learning
- Obstacles & Advanced Methods
  - Joint modeling
  - Collective inference
  - Scarcity of supervision signals
  - Diverse text genres and social media
- Remaining Challenges and Future Work
  - Rich semantic knowledge acquisition
  - Cross-lingual Wikification

References

http://nlp.cs.rpi.edu/kbp/2014/elreading.html

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials, page 7,
Baltimore, Maryland, USA, 22 June 2014.